Author Names & Affiliations

- Stephanie Teasley - University of Michigan
- Christopher Brooks - University of Michigan
- Abelardo Pardo - The University of Sydney
- Simon Buckingham Shum - Open University
- Xavier Oacha - Escuela Superior Politécnica del Litoral, Ecuador
- Dragan Gasevic - The University of Edinburgh

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Learning Analytics, education, learning sciences

Title of Submission

Educational Cyberinfrastructure

Abstract (maximum ~200 words).

Education and the Learning Sciences are now at the point of becoming data- and algorithm-intensive in the methodologies used to conduct research. In fact, education research is transitioning into a “Fourth Paradigm” discipline (Hey et al. 2009). The intersection of education with data science techniques has been a cornerstone of growth in research on learning over the last decade, and has translated into an unprecedented amount of data that needs to be securely collected, manipulated, and made available for analysis. Areas such as Multimodal Learning Analytics are producing dense streams of audio, video, psychophysiological and environmental data, and clickstreams that need to be managed by teams of stakeholders through infrastructure that guarantees its security and reliability, as well as student privacy. The application of analytic techniques to learning infrastructures has profound implications for how we both understand (at a theoretical level) and operationalize (at an institutional level) learning progression on all educational in professional learning. Therefore, Learning Analytics researchers require scalable infrastructure to pose the new kinds of questions that until recently were either unimaginable or impractical to investigate.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Education and the Learning Sciences are now at the point of becoming data- and algorithm-intensive in the methodologies used to conduct research. Education research is transitioning into a “Fourth Paradigm” discipline (Hey et al. 2009), and the intersection of education with data science techniques has been a cornerstone of growth over the last decade. The application of such techniques to learning infrastructures has profound implications for how we both understand (at a theoretical level) and operationalize (at an institutional level) learning progression on all educational in professional learning. This can be seen through the rapid growth of student success modeling systems, now a common feature in learning platforms (e.g. Blackboard, Canvas, Moodle) and educational administrative settings (e.g. Ellucian Signals, Civitas). Further, this has led to rapid scholarly investigation of data-driven approaches to understanding learning and the formation of new educational data science communities (including Educational Data Mining and Learning Analytics) with associated professional societies, conferences, journals, summer training institutes, and degree programs. The explosion of interest in “educational cyberinfrastructure” is reflected in the range of government educational reports (including U.S., European and Australasian) mapping the state of the art and future roadmaps (REFS).

RESEARCH QUESTIONS
Like many other communities entering this computing-intensive paradigm, Learning Analytics researchers require scalable infrastructure to pose questions that until recently were either unimaginable or impractical to investigate. Examples include:

- How do we provide personalized, real time feedback to learners at massive scale, based on the analysis of millions of permutations of activity traces from the use of digital tools such as simulations, design toolkits, and experimental apparatus? Who benefits from this feedback and how does it affect students’ behaviors?
- Given the capacity to aggregate data streams from myriad platforms, personal devices and environmental sensors, how can we design coherent visualizations to help educators and learners interpret and act productively?
- What are the most effective data models to aggregate different data streams and sources to provide meaningful and theoretically valid insights into learning?
- What are the ways in which human experts (e.g. teachers, instructors, mentors) will interpret and act on predictive models of student success which can be highly multidimensional and probabilistic in nature?
- The newest methods of text analytics can in principle codify a large corpus of student writing or online discourse automatically, performing in seconds what might take humans weeks of painstaking qualitative analysis. To develop such tools, researchers need to experiment systematically, with many iterations. How robust is the performance of automated analysis compared to human coding? Do analytic models transfer between learning contexts and institutions? What size training corpus is required for a return in performance? Can data be de-identified at scale in order to meet ethical requirements?
- What kind of data models most effectively represent the temporal nature of learning to assess progression and to provide contextualized feedback to learners and instructors?

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

The increasing technology mediation in learning and education has translated into an unprecedented amount of data that needs to be securely collected, manipulated, and made available for analysis. The overall procedure requires a flexible and yet extremely secure data infrastructure accompanied by a software layer to make the data available for applications. Areas such as Multimodal Learning Analytics are producing dense streams of audio, video, psychophysiological and environmental data, and clickstreams that need to be managed by teams of stakeholders through infrastructure that guarantees its security and reliability, as well as student privacy. The current approaches rely on off-the-shelf tools that have not been conceived for these experimental settings. In fact, such tools are suitable for scenarios that are significantly distant from those emerging in the context of educational research.

The infrastructure needed to address these challenges comprises:

- Secure, scalable platform with high communication bandwidth and storage capacity to store streams of data captured by systems such as video, audio, physiological and environmental data, and clickstreams.
- High performance data management procedures to execute analytical procedures (machine learning algorithms, computation of visualizations, etc.) in near real-time fashion (e.g. suitable for interventions in a classroom and at the individual student level).
- High availability services for data queries from external agents. In the current emerging ecosystems, the produced data streams need to be made available through high availability mechanisms (APIs, Web services) to other stakeholders to facilitate distributed analysis and translation into actionable knowledge, often joining data across modalities (e.g. clickstream data with student record data and in-classroom

audio data).
• A sociotechnical layer to the infrastructure which enables the fluid sharing and use of sensitive data (e.g. FERPA and COPPA protected) among legal entities (academic institutions, researchers, vendor partners) with ease.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

The upskilling of the user community is an often-neglected part of technology innovation programs, which typically results in new digital tools being used in rudimentary ways because it enables, or even requires, new ways of working that users are not ready for. In short, people change far more slowly than technology. Habits can be hard to change, especially among more senior educators and researchers with established ways of collaborating and conducting research. Thus, a coherent strategy for upskilling the workforce is required, arguably, focusing on the doctoral students and early career researchers who are strategically placed to introduce new practices into their research groups.

We propose that “educational cyberinfrastructure” can be harnessed to upskill the workforce needed to leverage cyberinfrastructure initiative. By “eating our own dog food”, therefore, we should deploy data-intensive tracking, analysis and feedback techniques to build the skills people will need to use those very same techniques in their own work.

We need to address questions such as:
• Can we design training platforms that can track the usage patterns of cyberinfrastructure tools, in order to understand usage and coach users to take them to the next level of expertise?
• Can we address the urgent educational data science skills shortage, and accelerate the development of the next generation of Learning Analytics researchers and educators, in order to instrument cyberinfrastructure tools, and to provide useful analytics in science and engineering educational platforms?

Consent Statement